Binary Journal of Data Mining & Networking 4 (2014) 49-53

http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Neural Network Approach In Water Quality Data Analysis For The River Narmada

Sanjeev Gour[1], Mamta Gour[2]

*Corresponding author:

Sanjeev Gour

[1]Research Scholar, Barkatullah University, Bhopal (M.P.)
[2]Govt. P.G.College, Harda (M.P.)

Abstract

Data Mining is the method used to find new, hidden, or unexpected patterns in data. In recent year, with the several availability of computing system with ever expanding capability, there is growing tendency to use data mining techniques to complement in data driven modeling, decision making and prediction. This study focuses on how data mining techniques are used for water quality Prediction is analyzed and also we have generate a classification and Prediction model with taking multilayer perception method of neural network.

Keywords: A Water quality, Data mining, classification model, Neural Network, Multilayer perceptron.

Introduction

Water is one of the most important needs in daily life which contains a major parts of earth's hydrosphere. There are many examples where data mining techniques are successfully being used in decision making and prediction, where large organizations and companies (business, marketing, medical, telecommunication, banks, infrastructural companies etc.) already benefit from data mining (Adriaans and Zantinge -1996, Fayyad et al -1996)[1,2]. This research presents the study of Neural Network classification technique in experimental data of river Narmada . Water quality is one of the major concerns of countries around the world. This study endeavors to automatically classify water quality. The water quality classes are evaluated using 6 factor indices. These factors are pH value (pH), Dissolved Oxygen.(DO), Biochemical Oxygen Demand (BOD), Nitrate Nitrogen (NO3N), Ammonia Nitrogen (NH3N) and PO4. The methodology involves applying data mining techniques using multilayer perception (MLP) neural network models. The data consisted of 20 years of Narmada River in district Harda M.P. (India).

Data Mining Concept

Data mining is an approach for information extraction from huge amount of data stored in a database (Miller and Han, 2001)[3]. Recent trends in information technology (IT) and its growing application areas in addition to increase of available databases, along with the data mining are being used to extract and interpret information available in the databases, and explore the necessary information and their relationships to produce useful information/knowledge for decision making.

Definition Translating Data mining word by word means the mining or digging in data with the purpose of finding information or respectively knowledge. Coming to the more abstract and very well known definition of Frawley, Data mining is defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley, 1992)[4].

Neural Network

Artificial neural networks (ANN) are a method in artificial intelligence. The idea is to simulate the structure and operation of biological classical neural networks. The structure consists of nodes, that can like a neuron, fire a signal or remain silent, according to a (in most cases) sigmoid activation function. Weighted edges transport the signals from one cell to another. Many topologies are possible, with or without hidden layers of cells or with or without feedback edges. For training, the edge weights are manipulated in order to reduce the training error. One common training strategy is back propagation network with two hidden units. The input data is transported to the cells in the input layer and the output is presented at the output layer. ANNs are instable classifiers because the performance heavily depends on the training data and are hard to interpret because the learnt model is coded in the edge weights. ANNs are used, among others, for pattern recognition purposes, especially image and speech recognition and, in bioinformatics for protein structure prediction.

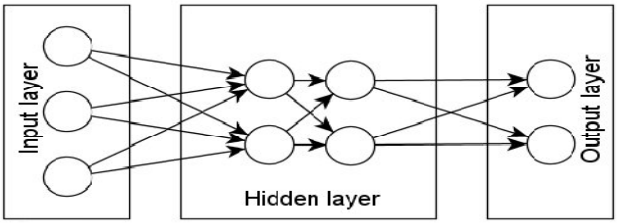

Figure 1- Simple artificial neural network. The network is a simple feed forward

Multilayer Perception Network

The artificial neural network (ANN) or neural network in short, is inspired by simulating the function of a human brain. A neural network can be used to represent a nonlinear mapping between input and output vectors. Neural networks are among the popular signal-processing technologies. In engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters. A general network consists of a layered architecture, an input layer, one or more hidden layers and an output layer.

Case Study Location

The data used in this study were collected at the river of Narmada. Samples are collected from the district of Madhya Pradesh name Harda, Monthly variations in the flow of river and the contributing drains were observed during monsoon and non-monsoon periods (1990-2012).

Experimental Setup

WEKA is an open source collection of machine learning algorithms for data mining tasks, and it includes neural network capabilities. First, we explore WEKA features and function by importing dataset samples, and trying to open the datasets with the software. Then we try to build ANN based on a particular dataset. The scope that covered by WEKA is wider than only ANN. We focus only on the ANN building functionality. Our experiment is limited to a pre-given dataset: (Water Quality Data of Harda District M.P from the analysis period 1990 to 2012 of Maa Narmada River).

List Of Parameters Selected

Many parameters can influence the surface water quality. Nine parameters are selected for the investigations after data pre-processing. The surface water quality can be classified as in following Table 1.

Table 1- List of water quality parameters

S.NO.	ATTRIBUTE	ABBREVIATION
1	PH	PH value
2	DO	Dissolve oxygen
3	BOD	Biochemical Oxygen Demand
4	No3_N	Nitrate Nitrogen
5	NH3_N	Ammonia Nitrogen
6	TEMP	Temperature
7	COD	Chemical oxygen demand
8	PO4	phosphate
9	Class (Polluted class)	A,B,C,D

Description Of Class Attribute

Generally, surface water quality can be divided into four classes;
class A - Extra clean fresh surface water resources use for conservation that are not necessary to pass through water treatment processes and require only ordinary processes for pathogenic destruction and ecosystem conservation where basic organisms can breed naturally;
class B -Very clean fresh surface water resources use for consumption that require ordinary water treatment processes before use by aquatic organisms in conservation, fisheries and recreation;
Class C- Medium clean fresh surface water resources use for consumption, but are passed through an ordinary treatment process before use;
Class D- Fairly clean fresh surface water resources use for consumption, but requires special water treatment processes before use.

Table 2- Class value for pollutants index

Pollutants index	Class			
	A	B	C	D
PH(Mg/l)	<5	5-9	5-9	5-9
DO(mg/l)	>6	6	4	2
BOD (mg/l)	<1.5	1.5	2	4
(No3/N)	<5	5	5	5
NH3N(mg/l)	<0.5	0.5	0.5	0.5

Experimental Results

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 3 -G -R
Relation:datarainyseassionofhardadis-weka.filters.unsupervised.attribute.Remove-R1
Instances: 23, Attributes: 9, parameters: pH_GEN_pH_units, Temp _C, BOD3, COD _mg_L, DO ,No3_N, O_PO4, NH3_N, pollution_Class
Test mode: evaluate on training data
Time taken to build model: 6.24 seconds

Table 3- Prediction of classes

Inst#	Actual	Predicted	Error prediction
1	1:A	1:A	0.903
2	2:B	3:C	+0.642
3	1:A	1:A	0.903
4	1:A	1:A	0.923
5	1:A	1:A	0.759
6	2:B	2:B	0.984
7	2:B	2:B	0.91
8	3:C	1:A	+0.887
9	2:B	2:B	0.836
10	3:C	3:C	0.986
11	1:A	1:A	0.925
12	1:A	1:A	0.763
13	2:B	2:B	0.787
14	2:B	2:B	0.923
15	2:B	2:B	0.817
16	2:B	2:B	0.789
17	3:C	1:A	+ 0.414
18	3:C	3:C	0.851
19	3:C	3:C	0.933
20	3:C	3:C	0.834
21	4:D	4:D	0.796
22	4:D	4:D	0.754
23	3:C	3:C	0.762

=== Predictions on training set ===

Table 4- MLP classifier performance criterion values

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.118	0.750	1.000	0.857	0.813	0.980	0.948	A
0.875	0.000	1.000	0.875	0.933	0.906	0.900	0.925	B
0.714	0.063	0.833	0.714	0.769	0.683	0.911	0.877	C
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	D

=== Evaluation on training set ===
Time taken to test model on training data: 0.02 seconds
=== Summary ===
=== Detailed Accuracy By Class ===

In this case, a confusion matrix is used as a basis for performance evaluation. The fields of the confusion matrix shown in Table 5, contain the numbers of examples of the following four subsets, Confusion matrix for classified sample data, representing a measure of accuracy in each class. While the training data for four classes show user accuracy of 100% and 87.5% and 71.42 % respectively.

Explanation of confusion matrix which has obtained for experiment:
Vertically reading from Target Class A, there are 6 records classified correctly. The accuracy percentage is 100%
Vertically reading from Target Class B, there are 7 records classified correctly. The accuracy percentage is 87.50%
Vertically reading from Target Class C, there are 5 records classified correctly. The accuracy percentage is 71.47%
Vertically reading from Target Class D, there are 2 records classified correctly. The accuracy percentage is 100%

Table 5- Confusion Matrix for experiment

Predicted Class	Target Class				
	A	6 100%	0	0	0
	B	1	7 87.5%	0	0
	C	2	0	5 71.42%	0
	D	0	0	0	2 100%
		A	B	C	D

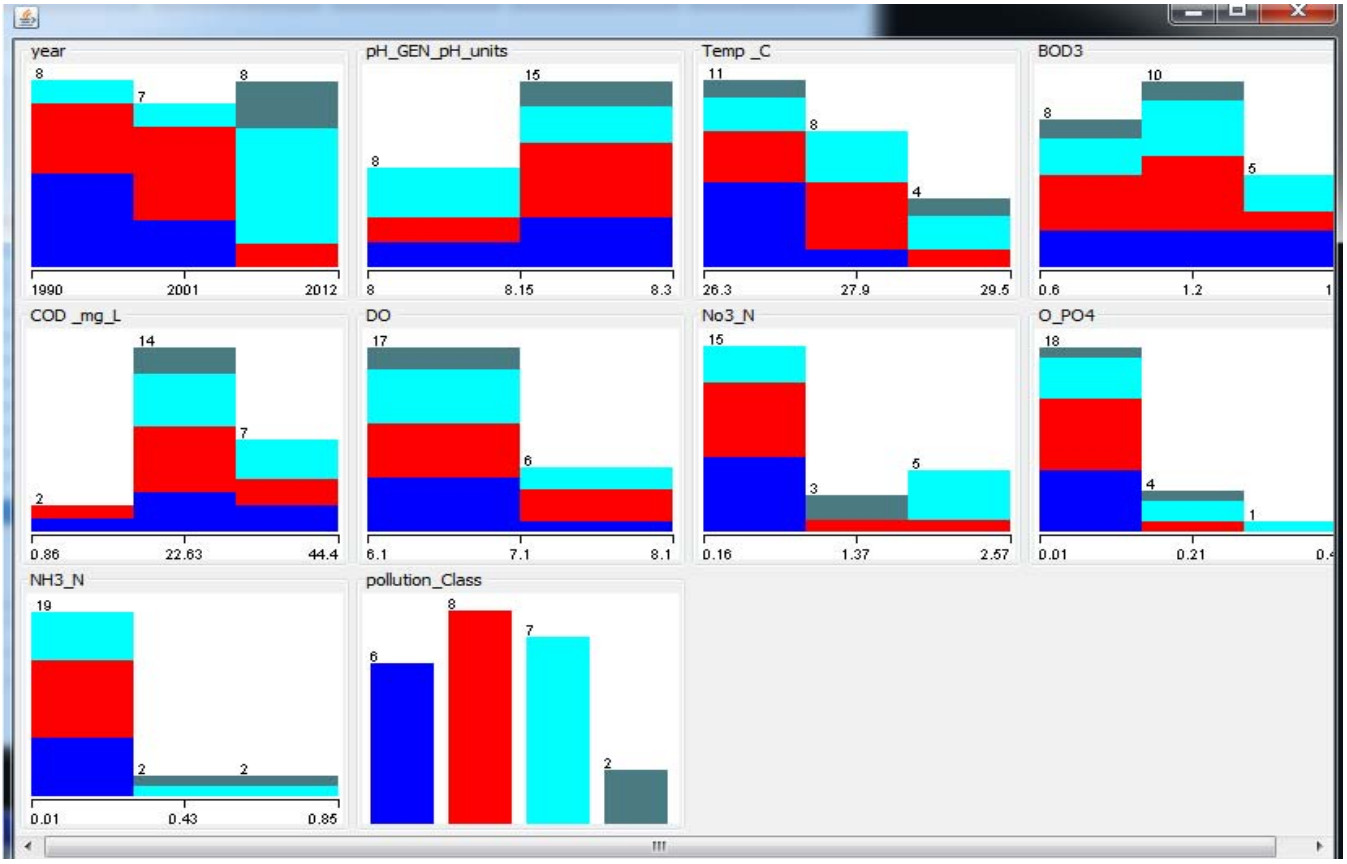

Figure 2- Attribute View With Weka

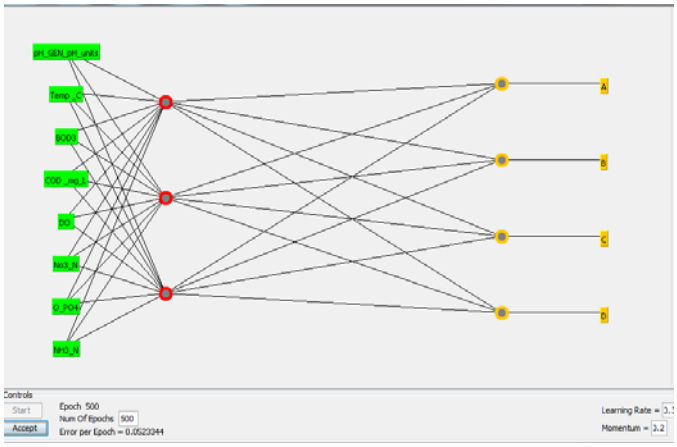

Figure 3- Neural network generated by Weka

All the training records were fed into the network to make it learn the potential relationships between water quality indices and their corresponding categories 4. *(See figure 3)*Accordingly, the input layer nodes represent 8 water quality indices, while the output layer nodes represent the 4 different class categories. The trained neural networks can provide an output representing the specific class for each of water quality indices. The testing samples are used to verify its classification ability. Many experimental investigations are conducted. The number of hidden nodes that provided the optimal result is 3 hidden nodes. The target error per epoch is 0.523344 in out of 100. Currently Instances classified 23; Attributes 9 .It can be seen that the network correctly classified 111 records from a total of 115 records. The accuracy percentage is 95.42 %. Optimal result is 3 hidden nodes. The target mean square error (MSE) is 0.001 after 500 iterations.

The neural network is using the given values of the 6 input variables to predict the class (A,B,C,D) So the training is to adjust the internal weights to get as close as possible to the known class values.

Conclusion

In this research we have used Multilayer perceptron neural network approach for prediction and classification of water quality parameters of river Narmada . Result of this chapter is summarized in table 3 to 5 and figure 2 and figure 3.

Some of interesting results are as follows:

From the output model of MLP classifier we have seen that parameters NH3_N and NO3_N have great influence on water

quality in different time and space. Also seen that the influence of BOD and DO with these nitrate composition causes the quality of water also improved at one level of class.So the results of MLP classifier indicated that the nitrate-nitrogen (NO3_N) of river water increased as the river traversed through the municipalities and industrial towns where effluents of deteriorated quality were added into the river.alsoas increase the value in NO3-N and DO in surface water, the quality of water also improved at one level of class.

Another interesting correlation found from the MLP classifier between BOD ,DO and pH value, We have found that the weight of attribute pH in every output class for class A is differ very largely from the threshold weight with respect to time (yearly) while DO and BOD have nearly similar weights it means there is no effect of these variable on the output class and the weight of attribute pH in every output class for class D is nearly similar from the threshold weight with respect to time (yearly) while DO and BOD have different weights. We have also seen that DO and BOD seem to have different weights and sign in all the neurons. So these are two highly opposite correlated variables. More water quality parameters should be modeled and their impact on the River should also be incorporated in the future research studies.

References

[1]. Adriaans P, and Zantinge D. Data mining, Addison-Wesley, 1996.

[2]. Fayyad UM. Piatetsky-Shapiro G, Smyth From Data Mining to Knowledge. P. (1996).

[3]. Han J, and Kamber M. Data Mining: Concepts and Techniques Morgan Kaufmann, 2001.

[4]. Frawley William J, Gregory Piatetsky-Shapiro, and Christopher J. Matheus "Knowledge Discovery in Databases: An Overview.AI Magazine Volume 13 Number 3 (1992).

[5]. Shoba G, Dr. Shobha G.(2014) " Water Quality Prediction Using Data Mining techniques"- International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 3 Issue 6 June, 2014 Page No. 6299-6306.

[6]. Nabeel M. Gazzaz , Mohd Kamil Yusoff, Ahmad Zaharin Aris , Hafizan Juahir(2012) "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors". Marine Pollution Bulletin.